Statistical Modelling XXXX; **XX**(**X**): 1–23

Semi-parametric Bayesian analysis of binary responses with a continuous covariate subject to non-random missingness

Frederico Z. Poleto[1], Carlos Daniel Paulino[2], Julio M. Singer[1] and Geert Molenberghs[3]
[1]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil
[2]Instituto Superior Técnico, Universidade Técnica de Lisboa (and CEAUL-FCUL), Av. Rovisco Pais, Lisboa, Portugal
[3]I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium, and Katholieke Universiteit Leuven, Leuven, Belgium

Abstract: Missingness in explanatory variables requires a model for the covariates even if the interest lies only in a model for the outcomes given the covariates. An incorrect specification of the models for the covariates or for the missingness mechanism may lead to biased inferences for the parameters of interest. Previously published articles either use semi-/non-parametric flexible distributions for the covariates and identify the model via a missing at random assumption, or employ parametric distributions for the covariates and allow a more general non-random missingness mechanism. We consider the analysis of binary responses, combining a missing not at random mechanism with a non-parametric model based on a Dirichlet process mixture for the continuous covariates. We illustrate the proposal with simulations and the analysis of a dataset.

Key words: Dirichlet process mixture; incomplete data; non-ignorable missingness mechanism; missing not at random; MNAR

Received October 2012; **Revised** December 2013; **Accepted** January 2014

1 Introduction

In many studies, data are missing for some explanatory variables ($\mathbf{X}$), and in order not to exclude either these sampling units or these variables from the analysis, we need to specify a model for their marginal distribution, or at least, for the conditional distribution of the explanatory variables that may be missing given the explanatory variables that are always observed even if the interest lies only on the conditional distribution of the response variables ($\mathbf{Y}$) given $\mathbf{X}$.

Address for correspondence: Frederico Z. Poleto, Instituto de Matemática e Estatística, Universidade de São Paulo, Caixa Postal 66281, São Paulo, SP, 05314-970, Brazil.
E-mail: frederico@poleto.com

 10.1177/1471082X14549290

When all variables in **X** are categorical and the number of combinations of their levels is much smaller than the number of sampling units, it may be reasonable to assume that **X** follows a multinomial distribution; this and other similar modelling strategies were studied by Ibrahim (1990), Gibbons and Hosmer (1991), Vach and Schumacher (1993), Lipsitz and Ibrahim (1996), Lipsitz, Parzen and Ewell (1998), Horton and Laird (1999), Satten and Carroll (2000) and Horton and Laird (2001) under assumptions of ignorable missingness. Analyses admitting non-ignorable missingness were proposed by Vach and Blettner (1995), Vach (1997), Ibrahim *et al.* (1999), Lipsitz *et al.* (1999) and Paik (2004).

In cases where at least one explanatory variable is continuous, we may not have a priori any information for a plausible parametric model. Hence, some authors adopt either semi-parametric or non-parametric models for the marginal distribution of **X** based on the assumption of ignorable missingness (Chen and Little, 1999; Chen, 2002, 2004, 2009; Zhang and Rockette, 2005, 2006, 2007; Zhao, 2009). Chen (2004) comments that his methodology can be extended to cases with non-ignorable missingness. Following an opposite stream, other authors consider parametric models for **X** along with non-random missingness mechanisms (Lipsitz *et at*. 1999; Stubbendick and Ibrahim, 2003, 2006; Huang *et al.* 2005; Miranda and Rabe-Hesketh, 2010). Ibrahim *et al.* (2005) present an excellent review on the analysis of generalized linear models with missing covariates. They cover not only the maximum likelihood and Bayesian approaches followed by most of the manuscripts referenced above, but also the popular multiple imputation technique (e.g., Rubin, 1987; Raghunathan *et al.*, 2001; Little and Rubin, 2002) and the relatively more recent weighted estimating equation methods (e.g., Robins *et al.,* 1994; Scharfstein and Irizarry, 2003).

Incorrect assumptions, either for the missingness mechanism or for the distribution of the covariates, may generate biased inferences for the conditional distribution of the responses given the covariates. We consider a flexible distribution for **X** along with sensitivity analyses for the missingness mechanism, allowing it to be non-random. We do not try to extend the the methodology proposed by Chen (2004) to the non-ignorable case because his approach is computationally intensive even in the case of ignorable missingness, and its repeated application required for the classical sensitivity analysis (Vansteelandt *et al.*, 2006) may be unfeasible. Therefore, we pragmatically adopt a Bayesian methodology for the sensitivity analysis of the missingness mechanism; this approach deals with the non-identifiability of the model through proper prior distributions. Furthermore, we use a non-parametric model for **X** based on a Dirichlet process mixture (Ishwaran and James, 2002). For simplicity, we restrict ourselves to the case of a single missing continuous covariate, although in Section 7 we comment on possible extensions to the multivariate case. A similar non-parametric Bayesian approach was employed by Scharfstein *et al.* (2003) for a single continuous response variable subject to missingness.

In Section 2, we describe the dataset that will be used to illustrate the methods considered in the remainder of the manuscript. In Section 3, we present an overview of the non-parametric Bayesian approach with Dirichlet process for complete data

analysis. In Section 4, we extend the model to additionally accommodate the missing data generating mechanism. We perform a simulation study to evaluate the proposed method in Section 5 and analyze our working example in Section 6.

2 The pulmonary embolism data

Wicki *et al.* (2001) analyzed data from 1090 patients that were consecutively admitted to the emergency ward of the University Hospital of Geneva for suspected pulmonary embolism, i.e., blockage of the main artery of the lung or one of its branches. We are interested in evaluating how the probability of occurrence of cardiovascular disease relates to diagnostic tests and other easily obtained information. For simplicity, we consider here only some of the explanatory variables included in the final model presented by these authors.

The indicator of the presence of pulmonary embolism (response variable) as well as four explanatory variables (age, previous pulmonary embolism or deep vein thrombosis, recent surgery, and pulse rate) were observed for all patients, while two variables that indicate presence of certain characteristics (platelike atelectasis and elevation of hemidiaphragm) had missing values for a single patient. On the other hand, the partial pressure of carbon dioxide ($PaCO_2$), obtained from arterial blood gas analysis, was missing for 103 (9%) patients.

In Figure 1, we exhibit histograms of $PaCO_2$ and Gaussian kernel method density estimates based on observed data and on sampled values from posterior predictive distributions obtained from the fit of the non-parametric model of Section 3 and from two parametric models in the normal, log-normal and gamma distribution families that had the best fit. The observed data seem to be better accommodated by the posterior predictive distribution of the non-parametric model than by the corresponding densities of the parametric models. The result is corroborated by Kolmogorov–Smirnov tests for the comparison of the empirical distribution of the observed data with the posterior predictive distributions for a new observation from the non-parametric, normal, log-normal and gamma models (p-values 0.207, 0.002, <0.001 and 0.004, respectively).

3 Non-parametric models for continuous variables with complete data

Let X_i, $i = 1, \ldots, n$, be a random sample of size n from a distribution function F. In the parametric approach, we assume a known form for F, indexed by a finite-dimensional parameter specified a priori, but generally unknown. To allow greater flexibility in modelling and robustness against misspecification of F, we consider non-parametric models; paradoxically this does not mean that the corresponding models are completely free from parameters, but rather indicates that the number and nature of parameters are variable and determined somehow by the data, potentially

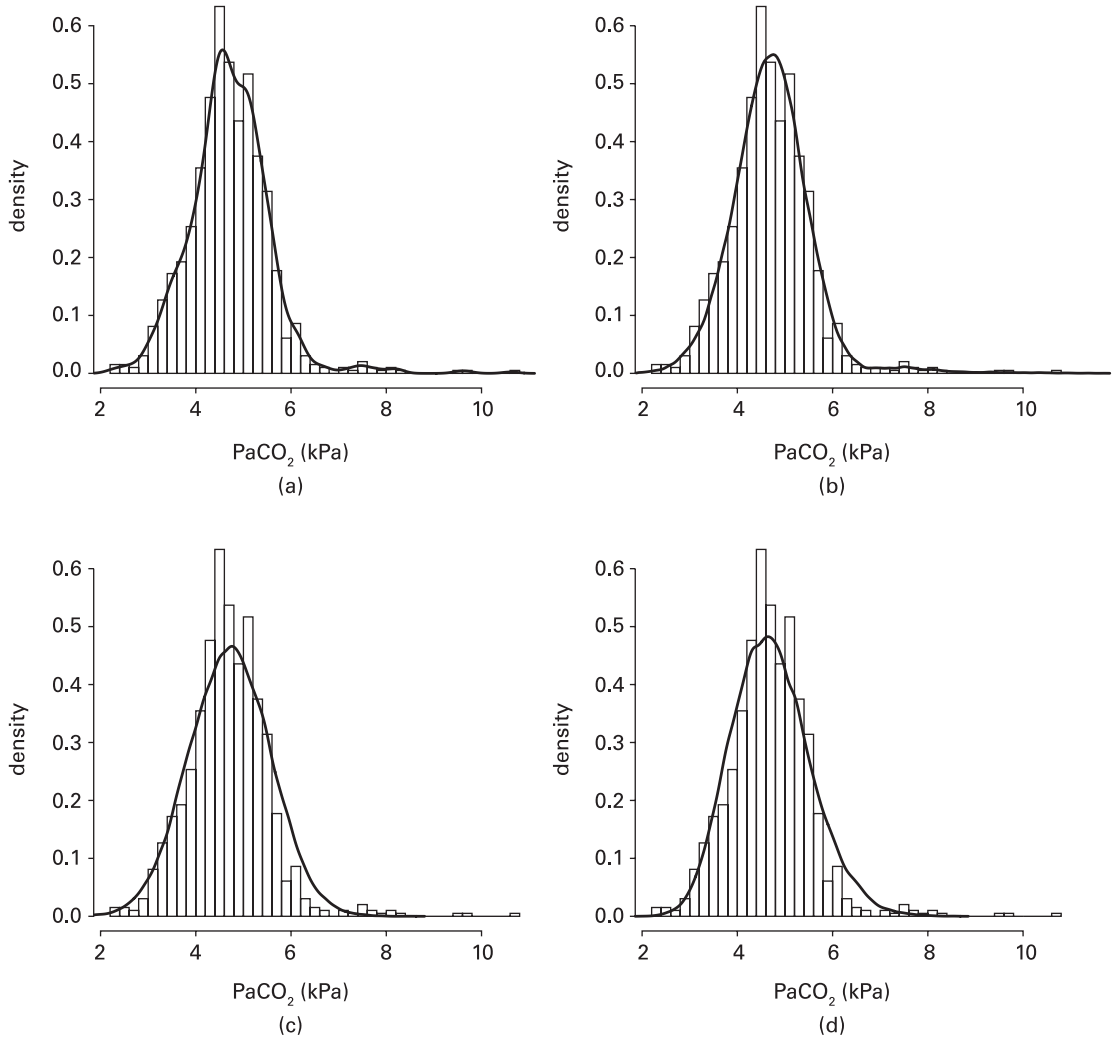

Figure 1 Histogram of the partial pressure of carbon dioxide ($PaCO_2$), in kPa, and density estimates obtained via the Gaussian kernel method based (a) on observed data and (b)-(d) on 50000 sampled values from the posterior predictive distributions obtained from the fit of (b) non-parametric, (c) normal and (d) gamma models.

reaching infinity. Reviews of some non-parametric methods in the Bayesian and classical paradigms are presented, respectively, by Müller and Quintana (2004) and Scott (1992).

One way to avoid the specification of the form of F is to employ random probability measures (RPM), which are probability distributions over the space of probability measures. Ferguson (1973) introduced the Dirichlet process (DP) as an RPM. Admitting that F follows a DP, symbolically, $F \sim \text{DP}(\alpha, F_0)$, means that for any

measurable partition $A_1, \ldots, A_M$ of the sample space, the probability vector $[F(A_1), \ldots, F(A_M)]$ follows a Dirichlet distribution with parameter vector $[\alpha F_0(A_1), \ldots, \alpha F_0(A_M)]$, where α is a precision parameter and F_0 is a reference distribution measured on the sample space. Under this parametrization, F_0 is the prior expectation of the distribution F and as α increases there is a greater concentration of F around F_0, up to the extreme case where $\alpha \to \infty$ indicates that F is assumed to be equal to F_0; on the other hand, small values of α (e.g., < 5) allow, in general, F to deviate considerably from F_0 (Congdon, 2006, p. 201). Given n independent and identically distributed observations, the posterior distribution is $F|(x_1, \ldots, x_n) \sim \text{DP}(\alpha + n, F_1)$, where $F_1 = (\alpha F_0 + nF_n)/(\alpha + n)$ and F_n is the empirical distribution function of the observations.

The simplicity of the properties of the DP and the ease with which the posterior distribution is obtained highlight why this model is so attractive. However, the DP generates a discrete distribution almost surely, which may not be appropriate for many applications. A simple way to generate an RPM compatible with absolutely continuous distributions is to assume that X_i follows an absolutely continuous distribution given the value of a specific parameter θ_i and that, in turn, θ_i, $i = 1, \ldots, n$, are a random sample of a DP, i.e.,

$$X_i \mid \theta_i \overset{\text{ind.}}{\sim} F_{\theta_i}, i = 1, \ldots, n, \tag{3.1}$$

$$(\theta_1, \ldots, \theta_n) \mid G \overset{\text{i.i.d.}}{\sim} G, \quad G \mid (\alpha, G_0) \sim \text{DP}(\alpha, G_0). \tag{3.2}$$

Assuming that the parameters $\{\theta_i\}$ follow a prior distribution of the DP type centred on G_0, instead of the common approach of assuming that these parameters directly follow a parametric distribution G_0 (Walker *et al*., 1999) adds the desired flexibility to the model. The term Dirichlet process mixture (DPM) stems from the hierarchical formulation (3.1)-(3.2) which implies that the marginal distribution for X_i is a mixture, i.e.,

$$f(x_i) = \int f(x_i \mid \theta_i)\, dG(\theta_i), \quad G \mid (\alpha, G_0) \sim \text{DP}(\alpha, G_0). \tag{3.3}$$

This model shall not be confounded with the mixture of Dirichlet processes (MDP), suggested by Antoniak (1974), a parametric mixture induced by distributions imposed on the parameters of the DP, regarded as hyper-parameters. As with the DP, the MDP generates a discrete distribution almost surely, whereas the DPM, considering or not prior distributions for the parameters of the DP, produces absolutely continuous distributions.

The constructive definition of the DP, presented by Sethuraman (1994), shows that $G|(\alpha, G_0) \sim \text{DP}(\alpha, G_0)$ can be represented by:

$$G(A) = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}(A) \tag{3.4}$$

for any measurable subset A of the space of values of $\{\theta_j\}$, where,

$$p_1 = V_1, \quad p_j = V_j \prod_{k=1}^{j-1} (1 - V_k), j > 1, \quad V_j \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha), \; j = 1, 2, \ldots, \tag{3.5}$$

$\delta_{\theta_i}(A)$ is the Dirac measure, i.e., is equal to one if $\theta_j \in A$ or to zero, otherwise, and

$$\theta_j \overset{\text{i.i.d.}}{\sim} G_0, \; j = 1, 2, \ldots \tag{3.6}$$

(Walker *et al.*, 1999). This very useful result allows the design of efficient algorithms to fit DPM models by rewriting (3.3) as:

$$f(x_i) = \sum_{j=1}^{\infty} p_j f(x_i \mid \theta_j), \; i = 1, \ldots, n. \tag{3.7}$$

The construction of the random weights $\{p_j\}$ for the mixture in (3.5) is the so-called stick-breaking procedure. Note that $\sum_{j=1}^{\infty} p_j = 1$.

In practice, however, for simplicity, it is common to truncate the mixture (3.7) to M components (see, e.g., Ishwaran and James, 2002), which is equivalent to approximating the DP(α, G_0) by a truncated Dirichlet process (TDP), denoted by TDP(α, G_0, M). In this case, to obtain the weights $p_1, \ldots, p_M$, we generate the variables $V_j \sim \text{Beta}(1, \alpha)$, $j = 1, \ldots, M - 1$, and set $V_M = 1$.

The choice of M is the key issue in the approach of obtaining TDP prior distributions. Firstly, we could appeal to the limit $M = n$, because at most one would have each sample unit x_i associated to a different θ_j in (3.7). Secondly, because the exact value of the continuous variable is often rounded by the measurement instrument or by the observation process, the number of distinct values contained in the sample is usually much smaller than n, and then, as in the previous case, it makes no sense to assume that M exceeds this number. In the study of pulmonary embolism, for example, from the 987 observed values of $PaCO_2$, there are only 243 distinct values. Finally, even if all values are distinct, Antoniak (1974) shows that (i) the DP naturally provides clusterings of $\{\theta_i\}$, assuming that nearby observations are associated to the same value of θ_j, (ii) a prior distribution for α induces a prior distribution for the number of distinct values of $\{\theta_i\}$, denoted by M^*, and (iii) for large n,

$$\text{E}(M^* \mid \alpha) \cong \alpha \ln\left(1 + \frac{n}{\alpha}\right). \tag{3.8}$$

In Table 1, we display some values for E($M^*|\alpha$), varying α and n in (8). The dependence of M^* and α can be understood from (5) because as α decreases, the distributions $\{V_j\}$ concentrate on values away from zero and therefore tend to have a smaller number of weights that are not as close to zero.

West (1992) and Escobar and West (1995) present analyses setting $\alpha = 1$, but

Table 1 Average number of distinct groups, $E(M^*|\alpha)$, obtained from (3.8), varying α and n.

$\alpha \backslash n$	1000	10000	30000	50000
2	12.4	17.0	19.2	20.3
3	17.4	24.3	27.6	29.2
5	26.5	38.0	43.5	46.1
8	38.7	57.1	65.8	69.9
10	46.2	69.1	80.1	85.2

both the latter along with Escobar and West (1998) and Ishwaran and James (2002), among others, suggest the use of a prior distribution given by:

$$\alpha|(\lambda_1, \lambda_2) \sim \text{Gamma}(\lambda_1, \lambda_2), \tag{3.9}$$

where $\text{Gamma}(\lambda_1, \lambda_2)$ denotes a gamma distribution with shape parameter λ_1 and scale parameter λ_2, such that the average is λ_1/λ_2. Following the suggestion of Ishwaran and James (2002), we consider $\lambda_1 = \lambda_2 = 2$, which concentrates around 98% of the α values between 0 and 3, allowing G to deviate considerably from G_0.

The posterior distribution of M^* can be used to evaluate the truncation of the DP, i.e., whether it was based on a too small value for M. Thus, if the 97.5% quantile of the posterior distribution of M^* is very close to M, it is reasonable to increase the value of M; on the other hand, if the 97.5% quantile of M^* is far from M, we can decrease the value of M without harming the approximation of the DP by the TDP and yet obtaining results faster, since larger values of M are associated with greater computational effort. The posterior distribution of α along with (3.8) may help in the selection of the new value. In the analysis of $PaCO_2$ in the pulmonary embolism study, for example, the 97.5% quantile of the posterior distribution of M^* was 11, a value reasonably smaller than the adopted $M = 20$; as the 97.5% quantile of the posterior distribution of α was a little smaller than 2, using the results in Table 1, we can decrease M to a value close to 13. The posterior distribution of α does not deviate much from the prior distribution, even with these substantial sample sizes, and this might be a consequence of some potential identifiability problems related to α as discussed by Leonard (1996).

One of the simplest DPM models is the Poisson-gamma, described by Escobar and West (1998) and Congdon (2006, p. 205), wherein F_{θ_i} in (3.1) is $\text{Poisson}(\theta_i)$ and G_0 in (3.2) is $\text{Gamma}(\lambda_1, \lambda_2)$. However, in the literature, a normal distribution is usually assumed for F. West (1992) explores connections between the DPM based on a normal distribution and density estimation techniques with Gaussian kernel. Escobar and West (1995) present one of the first developments that allowed an implementation of a Gibbs sampler for DPM, popularizing the approach. In these initial studies, the sampling schemes via Markov chain Monte Carlo (MCMC) integrate the DP, using its Pólya urn representation (Blackwell and MacQueen, 1973), since it is not possible to randomly select exact values from the DP. In our case, we do not integrate

the DP, but rather approximate it by the TDP, as discussed by Ishwaran and James (2002) and Congdon (2006, pp. 201–07). This is a pragmatic option, because the version of (3.7) truncated in M components can be easily implemented in software packages within the BUGS (Bayesian inference Using Gibbs Sampling) project—i.e., WinBUGS or OpenBUGS (Lunn *et al.*, 2000, 2009)—or JAGS (Just Another Gibbs Sampler, Plummer, 2003), and samples of the posterior distributions of interest can be obtained. In these computational approaches, it is common to introduce latent variables, s_i, that indicate which of the, θ_j, $j = 1, 2, \ldots$, is assigned to the i-th unit. This allows rewriting the truncated version of (3.7) as the hierarchical model:

$$
\begin{aligned}
X_i \mid \theta_{s_i} &\overset{\text{ind.}}{\sim} F_{\theta_{s_i}}, \quad i = 1, \ldots, n, \\
P(s_i = j) &= p_j, \ j = 1, \ldots, M, \ i = 1, \ldots, n, \\
p_1 = V_1, p_j &= V_j \prod_{k=1}^{j-1} (1 - V_k), \ j = 2, \ldots, M, \\
V_j &\overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha), \ j = 1, \ldots, M-1, \ V_M = 1, \\
\theta_j &\overset{\text{i.i.d.}}{\sim} G_0, \ j = 1, \ldots, M.
\end{aligned}
$$

Ishwaran and James (2001) show that this Gibbs sampler has the following characteristics.

1. The elements of $(\theta_1, \ldots, \theta_M)$ are conditionally independent given the other variables, and the full conditional distribution of θ_j is proportional to $g_0(\theta_j) \prod_{\{i: s_i = j\}} f(x_i \mid \theta_j)$, where g_0 is the density function of G_0.
2. The elements of $(V_1, \ldots, V_{M-1})$ are conditionally independent given the other variables, and the full conditional distribution of V_j is Beta$(1 + a_j, \alpha + b_j)$, where a_j is the number of $\{s_i\}$ equal to j and b_j is the number of $\{s_i\}$ greater than, j, $j = 1, \ldots, M - 1$.
3. The elements of $(s_1, \ldots, s_n)$ are conditionally independent given the other variables, and each s_i is updated from the distribution $P(s_i = j|x_i) \propto p_j f(x_i|\theta_j)$, $j = 1, \ldots, M$.

To avoid a higher degree of overparametrization of the model described in the next section, we consider the simplest version of the univariate normal model proposed by West (1992):

$$X_i \mid \mu_i, V \overset{\text{ind.}}{\sim} N(\mu_i, V), \ i = 1, \ldots, n, \tag{3.10}$$

$$(\mu_1, \ldots, \mu_n) \mid G \overset{\text{i.i.d.}}{\sim} G, \ G \mid (\alpha, G_0) \sim \text{DP}(\alpha, G_0), \tag{3.11}$$

where $N(\mu_i, V)$ denotes the normal distribution with mean μ_i and variance V, and

$$G_0 \mid (\mu_0, \tau, V) = N(\mu_0, \tau V). \tag{3.12}$$

West (1992) sets a value for μ_0 and adopts the prior distributions:

$$V^{-1}|(s_0, S_0) \sim \text{Gamma}(s_0/2, S_0/2), \tag{3.13}$$

$$\tau^{-1}|(w, W) \sim \text{Gamma}(w/2, W/2), \tag{3.14}$$

where S_0/s_0 is the prior guess for V and s_0 measures the prior belief in this guess. He did not mention how to choose w and W, but used $w = 2$ and $W = 10$ in his analysis, considering this an approximately diffuse distribution as a prior. Furthermore, he warns that, in general, there is not much information about τ in the sample, but that a prior for τ is necessary, paralleling the subjective choice of the smoothing parameter in kernel methods. Finally, he states that large values of τ induce a greater number of modes for the posterior predictive distribution of X. Escobar and West (1995) consider the heteroskedastic case, i.e., wherein each X_i in (3.10) may have a different variance, V_i, and adopt the prior

$$\mu_0|(a, A) \sim N(a, A), \tag{3.15}$$

with $A^{-1} \to 0$. In their example, they first evaluate the distribution of the number of modes induced by different values of τ and, then, use the hyper-parameters $w = 1$ and $W = 100$, considered compatible with their beliefs.

Ishwaran and James (2002) consider different alternatives. In particular, instead of (3.12), they assume,

$$G_0|(\mu_0, \tau) = N(\mu_0, \tau), \tag{3.16}$$

where the variance of the reference distribution for μ_i is a priori independent from the variance of X_i. Instead of using (3.14) as the prior distribution, they suggest to set the value of τ in such a way that (3.16) covers the values that $\{\mu_i\}$ may assume. They mention that a good choice is to take $\sqrt{\tau}$ as four times the standard deviation of the data. With respect to the prior distribution for V, they compare the results of (3.13) in the heteroskedastic case, using small values for s_0 and S_0 (e.g., $s_0 = S_0 = 0.02$), with those obtained under

$$V|T \sim \text{Unif}[0, T], \tag{3.17}$$

where Unif$[0, T]$ denotes a continuous uniform distribution in the range $[0, T]$, taking T equal to the variance of the data. They conclude that the gamma prior distribution can be too informative in certain circumstances, even when employing hyper-parameters related to non-informative distributions, and that this may smooth the data improperly. As a consequence, they suggest that the uniform prior distribution may be a more interesting alternative.

Due to the difficulty in choosing the hyper-parameters of (3.14) when using the reference distribution (3.12) and because it is possible that (3.13) may be more informative than assumed for the reference distribution (3.16), we follow the suggestions of Ishwaran and James (2002) adopting the uniform prior distribution for V. We do not consider here the heteroskedastic versions to avoid a higher degree of overparametrization, not only because the sample does not contain much information about α, but also because the more general missingness mechanisms considered in the next sections already suffer from identifiability problems.

4 A semi-parametric model for binary responses with a continuous covariate subject to non-random missingness

Let Y_i denote a binary response always observed, X_i, a continuous covariate with potentially missing values, and R_i, an indicator variable assuming the value of 1 if X_i is observed or 0, if X_i is missing, $i = 1, \ldots, n$. Although interest lies only in the conditional distribution of Y_i given X_i, it is necessary to consider a model for X_i, as we do not want to discard the portion of the sample wherein X_i is missing. As we admit that the missing data generating mechanism may depend on the unobserved values, we also need to model R_i.

Employing the so-called selection model factorization (Little and Rubin, (2002), we consider the model:

$$R_i \,|\, (Y_i, X_i, \delta_0, \delta_1, \delta_2, \delta_3) \overset{\text{ind.}}{\sim} \text{Bern}(\theta_i), \text{logit}(\theta_i) = \delta_0 + \delta_1 X_i + \delta_2 Y_i + \delta_3 X_i Y_i, \; i = 1, \ldots, n, \tag{4.1}$$

$$Y_i \,|\, (X_i, \beta_0, \beta_1) \overset{\text{ind.}}{\sim} \text{Bern}(\pi_i), \; \text{logit}(\pi_i) = \beta_0 + \beta_1 X_i, \; i = 1, \ldots, n, \tag{4.2}$$

$$X_i \,|\, (\mu_i, V) \overset{\text{ind.}}{\sim} N(\mu_i, V), \; i = 1, \ldots, n, \tag{4.3}$$

where Bern(θ_i) denotes the Bernoulli distribution with success probability θ_i, along with the prior distributions

$$\delta_j \,|\, (\mu_{\delta_j}, \sigma_{\delta_j}) \overset{\text{ind.}}{\sim} N(\mu_{\delta_j}, \sigma_{\delta_j}), \quad j = 0, 1, 2, 3, \tag{4.4}$$

$$\beta_j \,|\, (\mu_{\beta_j}, \sigma_{\beta_j}) \overset{\text{ind.}}{\sim} N(\mu_{\beta_j}, \sigma_{\beta_j}), \quad j = 0, 1, \tag{4.5}$$

$$(\mu_1, \ldots, \mu_n) \,|\, G \overset{\text{i.i.d.}}{\sim} G, \quad G \,|\, \alpha, G_0, M \sim \text{TDP}(\alpha, G_0, M), \tag{4.6}$$

$$V|T \sim \text{Unif}[0, T], \tag{4.7}$$

$$\alpha|(\lambda_1, \lambda_2) \sim \text{Gamma}(\lambda_1, \lambda_2), \tag{4.8}$$

$$G_0|(\mu_0, \tau) = N(\mu_0, \tau), \tag{4.9}$$

$$\mu_0|(a, A) \sim N(a, A), \tag{4.10}$$

all mutually independent.

Values for the hyper-parameters of these prior distributions are indicated in the applications. Note that model (4.1)–(4.3) does not lead to the likelihood of the observed data, but only to the likelihood of the complete data, which is precisely what is required by the computational packages of the BUGS project. Thus, in one of the stages of the MCMC sampling scheme, the algorithm randomly draws a value for the missing data from its conditional distribution given the other variables (observed and unobserved). Regarding the DP, in each iteration of the MCMC, the algorithm draws (1) μ_0, (2) μ_j, $j = 1, \ldots, M$, i.e., the M distinct values of $\{\mu_i\}$, (3) α, (4) V, (5) V_j, $j = 1, \ldots, M - 1$ (to obtain $p_1, \ldots, p_M$), and (6) chooses which of the μ_j, $j = 1, \ldots, M$, will be allocated to each μ_i, $i = 1, \ldots, n$.

The model is considered semi-parametric because it employs the non-parametric approach of the previous section for the marginal distribution of X_i and conventional parametric models for the conditional distributions of Y_i given X_i and R_i given Y_i and X_i.

The missingness mechanism (4.1) is non-random because it considers that the probability of having missing covariates may depend on their unobserved values. On the other hand, if we include the missing at random assumption

$$\text{MAR} : \delta_1 = \delta_3 = 0, \tag{4.11}$$

the missingness mechanism becomes ignorable under the viewpoint of Bayesian inferences for β_0 and β_1 due to the assumed prior independence between (δ_0, δ_2) and the other parameters (Little and Rubin, 2002). A subclass of the MAR model is the missing completely at random (MCAR) mechanism that can be formulated by setting

$$\text{MCAR} : \delta_1 = \delta_2 = \delta_3 = 0. \tag{4.12}$$

In this setup with missingness in explanatory variables, it is important to note that the so-called complete case analysis (CCA), where units with missing data are discarded, commonly generates unbiased inferences for β_0 and β_1 not only under the MCAR mechanism but also under any other missingness mechanisms that do not depend on the response Y_i such as in the reduced version of the missing not at random mechanism,

$$\text{MNAR}_{\text{red}} : \delta_2 = \delta_3 = 0. \tag{4.13}$$

A CCA of data generated under the non-random missingness mechanism (4.13) results in biased inferences for the marginal distribution of X_i, but not for the conditional distribution of Y_i given X_i. Also, the CCA does not require the specification of a marginal model for X_i if the interest lies only in the conditional distribution of Y_i given X_i.

5 Simulation study

We consider the following distributions for the explanatory variable

$$X^N \sim N(12, 3^2), \tag{5.1}$$

$$X^L \sim \text{Long-normal}(2.45, 0.246^2), \tag{5.2}$$

$$X^C = 0.8 \times X^{C1} + 0.2 \times X^{C2},\ X^{C1} \sim \text{Unif}[8, 12],\ X^{C2} \sim \text{Long-normal}(2.79, 0.642^2), \tag{5.3}$$

where Long-normal(μ, σ^2) denotes a log-normal distribution, and μ and σ are, respectively, the mean and the standard deviation of the underlying variable on the logarithmic scale. The mean and the standard deviation of X^L and X^C coincide with the corresponding parameters of X^N, although the densities are very different, as illustrated in Figure 2.

In order to assess the impact of results obtained under different distributional assumptions for the covariate, we generated 1 000 replicates of X from each of the three distributions (5.1), (5.2) and (5.3) with sizes $n = 50$, 100, 200, 400 and 1 000; then, for each value generated under each of the distributions of the covariates, we generated Y from (4.2) with $\beta_0 = 6$ and $\beta_1 = -0.5$; finally, we generated R from (4.1) with $\delta_0 = -3$, $\delta_1 = 0.5$ and $\delta_2 = \delta_3 = 0$. For each of the generated datasets (with X^N, X^L e X^C), we fitted the semi-parametric model of the previous section as well as

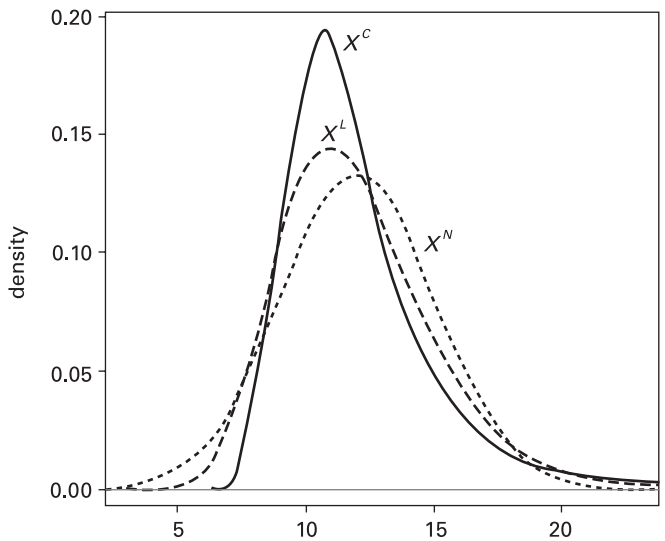

Figure 2 Densities of the distributions normal (X^N), log-normal (X^L) and linear combination (X^C) of an uniform and a log-normal.

normal and log-normal parametric models. For normal and log-normal parametric models, the non-parametric model (4.3) is replaced, respectively, by

$$X_i \mid (\mu_0, \tau) \overset{\text{i.i.d.}}{\sim} N(\mu_0, \tau), \quad i = 1, \ldots, n, \tag{5.4}$$

$$X_i \mid (\mu_0, \tau) \overset{\text{i.i.d.}}{\sim} \text{Log-normal}(\mu_0, \tau), \quad i = 1, \ldots, n, \tag{5.5}$$

and, for both, the prior distributions (4.6)-(4.9) are replaced by (14), with hyper-parameters associated with vague distributions, namely, large A (10^6 and 10^3 for, respectively, normal and log-normal cases) and small w (2 for both cases). For the semi-parametric model, the hyper-parameters of (4.6)-(4.10) were chosen as described in Section 3. For all models, we adopted vague prior distributions for δ_j and β_j employing the hyper-parameters $\mu_{\delta_j} = \mu_{\beta_j} = 0$ and $\sigma_{\delta_j} = \sigma_{\beta_j} = 10^3$, $j = 0, 1$. A full summary of the hyper-parameters is described in the Appendix. We always assumed the correct structure for the missingness mechanism, i.e., $\delta_2 = \delta_3 = 0$, so that the only varying components in the study are the distribution employed to generate the covariate and the distribution adopted for the covariate in the analysis. By applying standard diagnostic methods to evaluate the convergence of the Markov chains (Heidelberger and Welch, 1983; Gelman and Rubin, 1992; Geweke, 1992; Raftery and Lewis, 1992) generated for β_0 and β_1 on some of the analyses, we decided to generate 5 000 values for the burn-in of all Monte Carlo replicates, and then an additional 50 000 values for the chains, where a thinning interval of 10 values was finally used.

In Table 2, we display the Monte Carlo estimates for the coverage of the 95% equal-tailed credible intervals for β_0 and β_1 under CCA and analyses of all available data with parametric (normal and log-normal) and non-parametric models assumed for the covariate. An advantage of the MNAR model defined by constraint (4.13) is that we can compare the results to those obtained under the proposed models to that of the CCA, since in this setup it does not lead to biased inferences for β_0 and β_1. Consequently, the coverages for the CCA and for the correct parametric models were the closest to the desired level for all sample sizes. For small sample sizes, there were instances where incorrect parametric models exhibited coverages slightly closer to 95% than the ones of the non-parametric model, but in general, as sample sizes increased, the non-parametric model had results much better than the incorrect parametric models, even though for X^C the coverage was still far away from the desired level with $n = 1\,000$.

As the MNAR model under the constraint (4.13) is identifiable, we avoid dealing with problems caused by non-identifiability, at least initially. In order to explore this issue, we repeated the simulation study, i.e., we employed model (4.1) without considering constraint (4.13) by setting $\delta_0 = -6$, $\delta_1 = 0.5$, $\delta_2 = 1$ and $\delta_3 = 0.5$ to generate values for R. When fitting non-identifiable models for incomplete categorical data, Poleto *et al.* (2011) note that to assess convergence, chains much larger than the ones that would be required for identifiable models (such as a MAR model) are needed; they also observe that this scenario is further aggravated when larger samples

Table 2 Monte Carlo estimates for the coverage of the 95% equal-tailed credible intervals (in percentage) for β_0 and β_1 under CCA and analyses of all available data with parametric (normal and log-normal) and non-parametric models assumed for the covariate generated under an identifiable MNAR mechanism.

Covariate Distribution		$n = 50$		100		200		400		1 000	
Generated	Assumed	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
	No (CCA)	93	93	93	94	96	95	96	96	96	96
X^N	Normal	93	93	93	94	95	95	95	96	95	95
	Log-normal	93	92	93	92	95	94	94	94	92	93
	Non-parametric	87	86	88	88	92	91	95	96	95	95
	No (CCA)	92	92	93	93	94	94	94	94	95	94
X^L	Normal	91	91	89	90	86	87	72	75	44	49
	Log-normal	93	93	94	94	96	96	95	95	94	94
	Non-parametric	88	90	89	90	92	93	89	90	89	90
	No (CCA)	93	93	94	94	95	95	95	96	95	94
X^C	Normal	92	92	86	88	71	73	50	54	15	18
	Log-normal	94	94	92	93	84	86	76	79	49	54
	Non-parametric	92	92	91	92	85	86	81	84	78	80

or more vague prior distributions are considered. Therefore, we chose to use more concentrated prior distributions for δ_j, $j = 1, 2, 3$, i.e., with $\sigma_{\delta_j} = 1$. It is worth noting that Scharfstein *et al.* (2003) are, to our knowledge, the only authors to consider Dirichlet processes in the analysis of continuous incomplete data (although in the context of MDP and not DPM). They explore two strands of analysis: (1) using a large value for the precision parameter of the DP ($\alpha = 10\,000$), which in practice is equivalent to adopting the parametric reference distribution, and considering vague prior distributions for the other parameters, (2) employing a small value for the precision parameter ($\alpha = 1$) and a prior distribution even more informative than those adopted here for the non-identifiable parameter of the missingness mechanism (i.e., a non-zero mean and a standard deviation equal to 0.25 for the prior normal distribution). Moreover, although they consider the approach suggested by Gelman and Rubin (1992) to assess convergence, they do not mention having applied this criterion in their analyses; instead, they employ an informal approach, graphically evaluating if the estimates for the posterior densities of the parameters of interest, obtained after setting different values for the non-identifiable parameter, appear to be close to those expected under normal distributions. These authors do not relate the choices of prior distributions to the lack of identifiability of the model and/or to the difficulty in assessing the convergence of the chains; their choices, however, are in line with the comments anticipated in this paragraph. To get an idea of the effect of the choice of the mean for these more concentrated prior distributions, we considered first $\mu_{\delta_j} = 0$ and then $\mu_{\delta_j} = 1$, for $j = 1, 2, 3$. Note that for both prior distributions, the true values of δ_j, $j = 1, 2, 3$, are located in regions with not too small values for the corresponding densities, although these values are generally different from the

Table 3 Monte Carlo estimates for the coverage of the 95% equal-tailed credible intervals (in percentage) for β_0 under CCA and analyses of all available data with parametric (normal and log-normal) and non-parametric models assumed for the covariate with hyper-parameters $\mu_{\delta_j} = 0$ (P0) and = 1 (P1), for $j = 1, 2, 3$ for generated under a non-identifiable MNAR mechanism.

Covariate Distribution		$n = 50$		100		200		400		1000	
Generated	Assumed	P0	P1	P0	P1	P0	P1	P0	P1	P0	P1
	No (CCA)	61		43		20		2		0	
X^N	Normal	84	90	93	92	93	93	95	96	93	95
	Log-normal	75	86	69	83	57	66	40	52	44	51
	Non-parametric	81	90	87	96	90	98	93	95	93	96
	No (CCA)	60		45		15		2		0	
X^L	Normal	82	90	85	90	78	84	61	71	18	24
	Log-normal	75	91	84	93	87	93	88	93	91	93
	Non-parametric	73	91	78	93	83	94	91	97	92	92
	No (CCA)	59		46		20		3		0	
X^C	Normal	74	82	58	67	32	37	6	6	0	0
	Log-normal	79	89	77	86	66	73	32	43	2	2
	Non-parametric	73	88	67	91	76	92	80	93	84	89

prior means. The objective was to include cases where the prior guess is not very far from the true value of the parameters, but also cannot hit the bull's-eye. The other hyper-parameters were chosen as described in the preceding paragraphs.

In Table 3, we present the results with both prior hyper-parameters, but only for β_0 as the results for β_1 were pretty similar. We observe that, although the results depend on the prior distributions, the impact is actually smaller than expected. As in the present case the CCA leads to biased inferences for β_0, it is not surprising that the corresponding coverages are the worst ones. The coverages of the correct parametric models and the non-parametric model were in general closest to the desired level, especially for $n \geq 200$. By comparing Tables 2 and 3, we note that the incorrect parametric models had much worse results in the latter than in the former. This is probably a consequence of the fact that the average percentages of missing data were approximately 9% in the former and 19% in the latter.

6 Analysis of the pulmonary embolism data

Among several models fitted to data of Section 2, the most interesting for illustrative purposes is specified by

$$R_i \mid (\delta_0, \delta_1, \delta_2, \mathrm{LN}_i, \mathrm{LP}_i) \overset{\text{ind.}}{\sim} \mathrm{Bern}(\theta_i), \mathrm{logit}(\theta_i) = \delta_0 + \delta_1 \mathrm{LN}_i + \delta_2 \mathrm{LP}_i, i = 1, \ldots, n, \quad (6.1)$$

$$Y_i \mid (\beta_0, \{X_{ji}, \beta_j, j = 1, \ldots, 7\}) \overset{\text{ind.}}{\sim} \mathrm{Bern}(\pi_i), \mathrm{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{7} \beta_j X_{ji}, i = 1, \ldots, n, \quad (6.2)$$

where Y_i is the indicator of pulmonary embolism, the explanatory variables X_{1i}, …, X_{7i} are, respectively: (i) an indicator of recent surgery, (ii) an indicator of previous pulmonary embolism or deep vein thrombosis, (iii) an indicator of platelike atelectasis on chest x-ray film, (iv) an indicator of elevation of a hemidiaphragm on chest x-ray film, (v) age, in decades, (vi) arterial pulse rate, in hundreds of beats per minute (bpm) and (vii) partial pressure of carbon dioxide ($PaCO_2$), in kPa, R_i is the indicator of observation of $PaCO_2$ (X_{7i}) and

$$\mathrm{LN}_i = \begin{cases} \mathrm{LC}_i, \text{ if } \mathrm{LC}_i < 0, \\ 0, \text{ otherwise}, \end{cases} \tag{6.3}$$

$$\mathrm{LP}_i = \begin{cases} \mathrm{LC}_i, \text{ if } \mathrm{LC}_i > 0, \\ 0, \text{ otherwise} \end{cases} \tag{6.4}$$

$$\mathrm{LC}_i = \mathrm{logit}(\pi_i) - \mathrm{logit}(\widehat{\pi}), \tag{6.5}$$

where $\widehat{\pi}$ is the estimated prevalence of pulmonary embolism in the hospital, assumed known, by plugging in the proportion of pulmonary embolism in the sample (27%). Note that as the previous pulmonary embolism is one of the covariates, effect estimates of the other covariates are also interpreted conditionally on the previous response, and not marginally.

Perrier (personal communication) stated that $PaCO_2$ was missing for some patients because the arterial blood gas analysis was not performed, as patients were not very sick or were so sick that they needed the administration of oxygen. There are no records of which of the two reasons is responsible for the missing $PaCO_2$ data. Perrier's comments suggest that it is reasonable to assume that the probability of observing $PaCO_2$ (θ_i): (i) is maximum for patients with probability of pulmonary embolism (π_i) close to the prevalence of pulmonary embolism (π) and (ii) decreases as the probability of pulmonary embolism is farther from the prevalence. With this in mind, the piecewise regression model in (6.1) allows the θ_i to decrease with different speeds as $\pi_i \to 0$ or as $\pi_i \to 1$ and, thus, we expect that $\delta_1 > 0$ and $\delta_2 < 0$. By including parameters of the measurement process in the missing data generating mechanism, the model being used here is of a shared-parameter kind (Molenberghs and Kenward 2007). The selection model framework of Section 4 can be applied with this change (and likewise can the pattern-mixture model) without any difficulty. The only patient for whom the two chest x-ray data were missing was not considered in the analyses.

As in the previous sections, we need to specify a model for X_{7i} in addition to (6.1) and (6.2). We could adopt a conditional model for X_{7i} given the other explanatory variables. However, preliminary analyses show that only the pulse rate and the occurrence of previous pulmonary embolism or deep vein thrombosis helped to explain the variability of $PaCO_2$, although, very weakly, since the coefficient of

Table 4 Posterior means, standard deviations (SD) and 95% equal-tailed credible intervals (CI).

	Non-parametric Model				Complete Case Analysis			
Parameters	Mean	SD	CI 95%		Mean	SD	CI 95%	
intercept	–2.476	0.642	[–3.727;	–1.192]	–2.585	0.672	[–3.901;	–1.273]
recent surgery	1.375	0.268	[0.852;	1.903]	1.512	0.293	[0.943;	2.086]
previous embolism	1.080	0.177	0.735;	1.429]	1.087	0.187	[0.724;	1.453]
x-ray - PA	0.706	0.187	0.339;	1.070]	0.732	0.200	[0.343;	1.126]
x-ray - EH	0.590	0.189	[0.221;	0.962]	0.591	0.201	[0.198;	0.990]
age (×10 years)	0.268	0.046	[0.179;	0.359]	0.288	0.048	[0.195;	0.384]
pulse rate (×100 bpm)	1.158	0.331	[0.508;	1.809]	1.221	0.352	[0.538;	1.914]
$PaCO_2$	–0.405	0.101	[–0.609;	–0.209]	–0.429	0.102	[–0.631;	–0.231]
δ_0	2.624	0.223	[2.200;	3.077]				
δ_1	0.482	0.210	[0.082;	0.914]				
δ_2	–0.112	0.250	[–0.587;	0.399]				

PA: platelike atelectasis, EH: elevation of hemidiaphragm.

determination for the linear model was only 1% and we did not observe any clear non-linear association in the corresponding scatter plots. Having this in mind, we adopted a marginal rather than a conditional model for X_{7i}. We considered normal, log-normal and gamma parametric as well as the non-parametric models. We used prior distributions as described in the previous sections; the means and variances of the adopted normal distributions for β_j, $j = 0, \ldots, 7$ and δ_j, $j = 0, 1, 2$ are all equal to, respectively, 0 and 10^3. We show the posterior summaries for $\{\beta_j\}$ and $\{\delta_j\}$ for the non-parametric model and the ones for $\{\beta_j\}$ for the complete case analysis in Table 4. As opposed to the simulation study, results for the normal, log-normal and gamma parametric models were pretty similar to the corresponding ones of the non-parametric model. In all analyses, the magnitudes of the Monte Carlo errors were smaller than the precision of the figures presented in the table. Standard diagnostic methods were used to evaluate the convergence of the Markov chains generated for β_0 and β_1 (Heidelberger and Welch, 1983; Gelman and Rubin, 1992; Raftery and Lewis, 1992) and did not show evidence against their convergence.

With the exception of the parameter associated with $PaCO_2$, the posterior standard deviations of the other parameters are in general smaller in the analyses that include all the data than in the complete cases analysis. In Figure 3, we display the estimates of θ_i (probability of observing $PaCO_2$) obtained from the posterior means of δ_0, δ_1 and δ_2 of the non-parametric model, and estimates of π_i (probability of pulmonary embolism), calculated for the observed data. The probability of observing $PaCO_2$ is higher for patients with high rather than low probability of pulmonary embolism. By using the information that the probability of observing $PaCO_2$ is smaller either for cases where the probability of pulmonary embolism is less or is greater than the prevalence, the model yields a weaker association between the presence of pulmonary embolism and the values of $PaCO_2$ than the corresponding one obtained in models

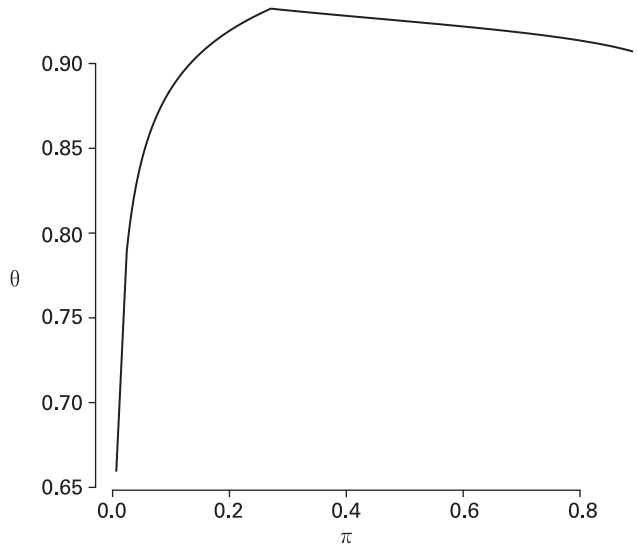

Figure 3 Estimates of π_i (probability of pulmonary embolism), obtained from the posterior means of $\{\beta_j\}$, and estimates of θ_i (probability of observing $PaCO_2$), obtained from the posterior means of $\{\delta_j\}$, calculated for the observed data.

that do not take these assumptions into account. Note, for example, the difference between the posterior means of $PaCO_2$ in Table 4. Analyses of all available data are more suitable than complete case analyses, because, by embedding assumptions about missing data, they should provide less biased results on the association between pulmonary embolism and $PaCO_2$, and generate more precise results for the other associations.

7 Discussion

We focused on the modelling of binary responses in the case where a single continuous explanatory variable has non-random missing values. We showed that the Bayesian approach with a non-parametric model based on a Dirichlet process mixture for the continuous covariate is a viable alternative to avoid possible biases in the inferences of interest introduced by the choice of an incorrect parametric distribution. In line with Poleto *et al.* (2011), Bayesian sensitivity analyses of the missingness mechanism via over-parameterized models and proper prior distributions, allows one to avoid too stringent untestable assumptions, while still leading to reasonable answers, i.e., interval estimates that, even though being wider due to the additional ignorance about the missingness mechanism, still contain the true values if the prior distributions embrace the correct missingness model.

Some additional extensions may be considered. Firstly, two or more continuous variables may have missing data. This must be considered both in the model for the missingness mechanism and in the model for the explanatory variables. Both can be specified with multivariate distributions or with a product of univariate conditional distributions. For the missingness mechanism, some authors (e.g., Lipsitz and Ibrahim, 1996; Ibrahim *et al*., 1999) express a preference for the latter strategy, that is, use of a product of Bernoulli distributions instead of a multivariate Bernoulli distribution. For the explanatory variables, we also believe that it may be more practical to work with unidimensional Dirichlet process mixtures, as described by Ishwaran and James (2002), than with the multivariate version of the non-parametric model considered by Escobar and West (1995), where a multivariate normal distribution with the usual multivariate normal and inverted Wishart prior distributions are employed. Furthermore, the modelling setting of Escobar and West (1995) is an extension of the univariate case described by West (1992) that, as discussed in Section 3, presents some difficulties with regard to the choice of hyper-parameters. On the other hand, by employing unidimensional Dirichlet process mixtures, the normal and uniform priors of Ishwaran and James (2002) can be used as long as the priors for the Dirichlet process mixtures can be considered all mutually independent.

Secondly, there might be interest in considering other distributions for the responses as well as multivariate cases. The replacement of these distributions may be performed in the modelling setting with nearly the same effort required in complete case analyses.

Thirdly, some of the response variables may be subject to missingness. These cases can be handled by a simultaneous modelling of the indicators of observation for these variables.

Finally, even though the Dirichlet process mixture is the non-parametric Bayesian approach most frequently employed in the literature, other non-/semi-parametric alternatives for the distribution of covariates can be considered, such as the Pólya tree prior distribution, a generalization of the Dirichlet process, resulting in a random probability measure compatible with continuous distributions. Paddock (2002) used this type of prior distribution on the analysis of responses with ignorable missingness. However, the use of this type of prior distribution generates predictive distributions with discontinuities, which may be inappropriate in some situations.

With or without the extensions described above, the biggest challenges for applying the models we deal with are likely the cases wherein assumptions for the missingness mechanism generate non-identifiable models and the sample size is too large and/or the prior distributions for all parameters are too vague. Under these conditions, the samples generated for the posterior distribution obtained by MCMC become extremely autocorrelated, thus, requiring very long chains to detect convergence and also to obtain Monte Carlo errors small enough to ensure the desired precision in the inferences of interest. In particular, these scenarios already highlighted in Poleto *et al.* (2011) become more severe with large sample sizes because the approximation of the Dirichlet process by its truncated version requires a greater number of components. Possibly other MCMC schemes as those described in Griffin and Holmes (2010) can attenuate this problem.

Acknowledgements

We gratefully acknowledge the financial supports to this research: Frederico Z. Poleto and Julio M. Singer, from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil (project 308613/2012-2); Carlos Daniel Paulino, from Fundação para a Ciência e Tecnologia (FCT) through the research centre CEAUL-FCUL, Portugal and project Pest-OE/MAT/UI0006/2014; Geert Molenberghs, from the IAP research Network P7/06 of the Belgian Government (Belgian Science Policy). The authors are grateful to Dr. Arnaud Perrier and to Dr. Henri Bounameaux, from Division of General Internal Medicine of Geneva University Hospitals, for providing the data.

Appendix: Hyper-parameter values

The hyper-parameters employed in Sections 5 and 6 are summarized in Table 5

Table 5 Parameter values for priors and hyper-priors

Parameter	Identifiable MNAR (Section 5)	Non-identifiable MNAR (Section 5)	Pulmonary Embolism (Section 6)
$\mu_{\delta j}$, $j = 1, \ldots$	0	*	0
$\sigma_{\delta j}$, $j = 1, \ldots$	1,000	1	1,000
$\mu_{\delta 0}$ and $\{\mu_{\beta j}\}$	0	0	0
$\sigma_{\delta 0}$ and $\{\mu_{\beta j}\}$	1,000	1,000	1,000
M	10	10	15
T	10	10	0.75
λ_1	2	2	2
λ_2	2	2	2
τ	160	160	12
a	0	0	0
A	1,000	1,000	1,000

*All 0 or all 1, depending on the hyper-parameters set.

References

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2**, 1152–74.

Blackwell D and MacQueen JB (1973) Ferguson distributions via polya urn schemes. *The Annals of Statistics*, **1**, 353–55.

Chen HY (2002) Double-semiparametric method for missing covariates in Cox regression models. *Journal of the American Statistical Association*, **97**, 565–75.

Chen HY (2004) Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the*

American Statistical Association, **99**, 1176–89.

Chen HY (2009) Estimation and inference based on neumann series approximation to locally efficient score in missing data problems. *Scandinavian Journal of Statistics*, **36**, 713–34.

Chen HY and Little RJA (1999) Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, **94**, 896–908.

Congdon P (2006) *Bayesian Statistical Modelling*, 2nd ed. New York: Wiley.

Escobar MD and West M (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–88.

Escobar MD and West M (1998) Computing nonparametric hierarchical models. In Dey D, Müller P and Sinha D, eds. *Practical nonparametric and semiparametric Bayesian statistics (Lecture notes in statististics 133)*. New York: Springer, 1–22.

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–30.

Gelman A and Rubin DB (1992) A single series from the gibbs sampler provides a false sense of security. In Bernardo JM, Berger JO, Dawid, AP and Smith AFM, eds. *Bayesian Statistics 4*. Oxford: Oxford University Press, 625–31.

Geweke J (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In Bernardo JM, Berger JO, Dawid AP and Smith AFM, eds., *Bayesian Statistics 4*. Oxford: Oxford University Press, 169–93.

Gibbons LE and Hosmer DW (1991) Conditional logistic regression with missing data. *Communications in Statistics—Simulation and Computation*, **20**, 109–20.

Griffin J and Holmes C (2010) Computational issues arising in Bayesian nonparametric hierarchical models. In Hjort NL, Holmes C, Müller P and Walker SG, eds. *Bayesian Nonparametrics*. Cambridge: Cambridge University Press, 208–22.

Heidelberger P and Welch PD (1983) Simulation run length control in the presence of an initial transient. *Operations Research*, **31**, 1109–44.

Horton NJ and Laird NM (1999) Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, **8**, 37–50.

Horton NJ and Laird NM (2001) Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics*, **57**, 34–42.

Huang L, Chen M-H and Ibrahim JG (2005) Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics*, **61**, 767–80.

Ibrahim JG (1990) Incomplete data in generalized linear models. *Journal of the American Statistical Association*, **85**, 765–69.

Ibrahim JG, Chen M-H, Lipsitz SR and Herring AH (2005) Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*, **100**, 332–46.

Ibrahim JG, Lipsitz SR and Chen M-H (1999) Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **61**, 173–90.

Ishwaran H and James LF (2001) Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–73.

Ishwaran H and James LF (2002) Approximate Dirichlet process computing finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, **11**, 508–32.

Leonard T (1996) On exchangeable sampling distributions for uncontrolled data. *Statistics & Probability Letters*, **26**, 1–6.

Lipsitz SR and Ibrahim JG (1996) A conditional model for incomplete covariates in

parametric regression models. *Biometrika*, **83**, 916–22.

Lipsitz SR, Ibrahim JG, Chen M-H and Peterson H (1999) Non-ignorable missing covariates in generalized linear models. *Statistics in Medicine*, **18**, 2435–48.

Lipsitz SR, Parzen M and Ewell M (1998) Inference using conditional logistic regression with missing covariates. *Biometrics*, 54, 295–303.

Little RJA and Rubin DB (2002) *Statistical analysis with missing data*, 2nd ed. New York: John Wiley & Sons.

Lunn DJ, Spiegelhalter D, Thomas A and Best N (2009) The BUGS project: evolution, critique and future directions (with discussion). *Statistics in Medicine*, **28**, 3049–82.

Lunn DJ, Thomas A, Best N and Spiegelhalter D (2000) WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–37.

Miranda A and Rabe-Hesketh S (2010) Missing ordinal covariates with informative selection. Tech. rep., NCRM Working Paper. Department of Quantitative Social Science, Institute of Education, University of London.

Molenberghs G and Kenward MG (2007) *Missing data in clinical studies*. New York: Wiley.

Müller P and Quintana FA (2004) Nonparametric Bayesian data analysis. *Statistical Science*, **19**, 95–110.

Paddock SM (2002) Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse. *Biometrika*, **89**, 529–38.

Paik MC (2004) Nonignorable missingness in matched case-control data analyses. *Biometrics*, **60**, 306–14.

Plummer M (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 20–22.

Poleto FZ, Paulino CD, Molenberghs G and Singer JM (2011) Inferential implications of over-parameterization: a case study in incomplete categorical data. *International Statistical Review*, **79**, 92–113.

Raftery AE and Lewis SM (1992) How many iterations in the Gibbs sampler? In JM Bernardo, JO Berger, AP Dawid, and AFM Smith, eds, *Bayesian Statistics 4*. Oxford: Oxford University Press, 763–73.

Raghunathan TE, Lepkowski JM, Van Hoewyk J and Solenberger P (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey *Methodology*, **27**, 85–95.

Robins JM, Rotnitzky A and Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846–66.

Rubin DB (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Satten GA and Carroll RJ (2000) Conditional and unconditional categorical regression models with missing covariates. *Biometrics*, **56**, 384–88.

Scharfstein DO, Daniels MJ and Robins JM (2003) Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics*, **4**, 495–512.

Scharfstein DO and Irizarry RA (2003) Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. *Biometrics*, **59**, 601–13.

Scott DW (1992) *Multivariate density estimation: theory, practice, and visualization*. New York: Wiley.

Sethuraman J (1994) A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–50.

Stubbendick AL and Ibrahim JG (2003) Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, **59**, 1140–50.

Stubbendick AL and Ibrahim JG (2006) Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, 16, 1143–67.

Vach W (1997) Some issues in estimating the effect of prognostic factors from incomplete covariate data. *Statistics in Medicine*, **16**, 57–72.

Vach W and Blettner M (1995) Logistic regression with incompletely observed categorical covariates—investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine*, **14**, 1315–29.

Vach W and Schumacher M (1993) Logistic regression with incompletely observed categorical covariates: a comparison of three approaches. *Biometrika*, **80**, 353–62.

Vansteelandt S, Goetghebeur E, Kenward MG and Molenberghs G (2006) Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, **16**, 953–79.

Walker SG, Damien P, Laud PW and Smith AFM (1999) Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **61**, 485–27.

West M (1992) Modelling with mixtures (with discussion). In Bernardo JM, Berger JO, Dawid AP and Smith, eds., *Bayesian Statistics 4*. Oxford: Oxford University Press, 503–24.

Wicki J, Perneger TV, Junod AF, Bounameaux H and Perrier A (2001) Assessing clinical probability of pulmonary embolism in the emergency ward. *Archives of Internal Medicine*, **161**, 92–97.

Zhang Z and Rockette HE (2005) On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference*, **134**, 206–23.

Zhang Z and Rockette HE (2006) Semiparametric maximum likelihood for missing covariates in parametric regression. *Annals of the Institute of Statistical Mathematics*, **58**, 687–706.

Zhang Z and Rockette HE (2007) An EM algorithm for regression analysis with incomplete covariate information. *Journal of Statistical Computation and Simulation*, **77**, 163–73.

Zhao Y (2009) Regression analysis with covariates missing at random: a piece-wise nonparametric model for missing covariates. *Communications in Statistics—Theory and Methods*, **38**, 3736–44.